HACIENDO ESTUDIOS ESTADÍSTICOS

Lee esta nota de prensa publicada por el **INE** (Instituto Nacional de Estadística, 2006):

Sedentarismo, hábitos alimenticios y obesidad

El 60,6% de la población de 16 y más años (63,6% de los hombres y 57,6% de las mujeres) realiza actividad física en su tiempo libre. Este porcentaje alcanza el 80,3% en el grupo de población infantil, aunque un 17,6% de los niños y un 21,9% de las niñas son sedentarios.

En cuanto a los hábitos de alimentación, un 13,4% de la población de uno y más años realiza un desayuno completo (lácteo u otro líquido más fruta o zumos más hidratos de carbono), mientras que el 2,9% no desayuna nada.

El 11,2% de la población (9,3% de los hombres y 13,0% de las mujeres) sigue una dieta o régimen especial. De ellos, el 49,9% la sigue debido a problemas de salud.

El 37,8% de las personas de 18 y más años tiene sobrepeso y un 15,6% presenta obesidad.

Entre la población de dos a 17 años, el 18,7% tiene sobrepeso y el 8,9% es obeso. Tanto en hombres como en mujeres, la obesidad es más frecuente a mayor edad, (excepto en los mayores de 74 años).

Ahora plantéate lo siguiente:

- ¿Has entendido la información?
- ¿Cómo sabe el INE todo eso si a ti ni a mí nos han preguntado? ¿A quiénes preguntan?
- ¿Qué cuentas han hecho los del INE para saber toda esa información?

En este tema **aprenderás cosas relacionadas con la estadística**, desde cómo se hace una **encuesta** hasta cómo hacer **cuentas** con los **datos** y sacar **conclusiones**, todo ello sobre hábitos de alimentación saludable.

¿Te parece interesante?... Pues entonces, adelante, que la puerta está abierta...

¿Cómo se empieza un estudio estadístico como el de la nota de prensa?

Debemos tener en cuenta:

- ¿Qué queremos saber?
- ¿De quién queremos saber la información?
- ¿Cómo obtenemos los datos?

Lo primero es decidir qué quiero saber

Por ejemplo, según la noticia que has leído, un 2,9% de los españoles no desayuna nada. Sin embargo el desayuno es algo fundamental para nuestra salud.

Nosotros **queremos estudiar el tiempo dedicado a desayunar**, ya que es difícil hacer un desayuno adecuado con prisas.

Empezando a adquirir vocabulario

A la característica o cualidad que queremos estudiar la llamaremos variable estadística.

Otros ejemplos de variable estadística

- El "tiempo dedicado a desayunar" sería una variable estadística. Te ofrecemos otros ejemplos:
- La marca de cereales para el desayuno.
- El número de horas de sueño.
- El color de los ojos.
- La estatura.
- El número de libros leídos el último mes.

Pero **no todas las variables estadísticas son iguales**. Las hay básicamente de dos clases, según el tipo de datos que estudiemos:

Variables cuantitativas: son aquellas que pueden medirse numéricamente. Por ejemplo, "*el tiempo dedicado a desayunar*", que puede expresarse en número de minutos).

Dentro de las variables cuantitativas las hay continuas y discretas: ¿Qué diferencia hay entre medir el número de veces que se comen legumbres a la semana y el peso? ¿a que uno no come 3,4567 veces legumbres? ¡O come 3 veces o 4! Pero sí podemos pesar 69,658 kg.

- **Variable discreta:** tiene un número FINITO de posibles resultados ("paramos de contar" las posibilidades).
- **Variable continua:** el número de posibles respuestas es infinito, así que debemos **agruparlas en intervalos**. Por ejemplo, como son demasiadas las posibles respuestas si preguntamos el peso, agrupamos las respuestas en intervalos: "entre 50 y 60 kilogramos", "más de 60 y hasta 70 kg", etc. Puede que aunque no haya infinitas posibilidades sean demasiadas y también debamos agruparlas.

Variables cualitativas: no todas las cosas que podríamos estudiar estadísticamente pueden expresarse con números. En algunos casos se trata de cualidades no medibles numéricamente. Por ejemplo, "*la marca de cereales para el desayuno*" se expresará de forma no numérica, con el nombre de la marca ("Estadifibra", "Matechoco", etc).

Recuerda:

Hay variables estadísticas cuantitativas (que pueden ser discretas o continuas) y cualitativas

Comprueba que lo has entendido

1. De la lista de variables que has visto en el ejemplo, señala las que son cualitativas e indica cuáles de ellas son continuas.
2. Escribe un ejemplo, distinto a los vistos, de variable cualitativa y otro de variable cuantitativa. Pero que estén relacionadas con el desayuno.

¿De quién quiero saber la información?

Lo segundo que debemos decidir es **¿de quién o qué queremos saber la información?** Es decir ¿a qué personas vamos a preguntar o en qué objetos vamos a medir la característica que queremos estudiar?

- **El grupo de personas u objetos (individuos) en el que vamos a estudiar la variable estadística se llama población (P).**
- **El número de individuos de la población se llama tamaño poblacional (N).**

En nuestra nota de prensa. . .

En el caso de nuestra nota de prensa, la población P es el conjunto de todos los españoles y N es el número de habitantes de España.

Pero... ¿A que a ti no te preguntó nada nadie del INE? Ni a nosotros tampoco. Entonces... ¿cómo se obtuvo la información? El coste en tiempo, dinero y personal que supondría preguntar a todos los españoles sobre sus hábitos sería gigantesco, por eso lo que se hace es **seleccionar una muestra**.

- **Una muestra es una parte de la población sobre la que estudiaremos la variable estadística.**
- **El número de individuos de la muestra es el tamaño muestral, n.**

El objetivo es **extender las conclusiones que se obtengan sobre la muestra a TODA la población**. Así, preguntando a un grupo de españoles en vez de a todos, el INE extrae conclusiones sobre la población española total (y se ahorra una pasta...).

Pero pensemos un poco...

Comprueba que lo has entendido

3. ¿Deberíamos tener cuidado con el número de individuos de la muestra?
4. ¿Influirán las características de los individuos de la muestra en las conclusiones finales?

La selección de una muestra adecuada es fundamental si queremos que lo que estudiemos en ella pueda extenderse a toda la población.

¡Cuidado!

La muestra debe ser representativa.

Los estadísticos disponen de técnicas adecuadas de selección de muestras y nosotros... también. Lo primero es distinguir entre:

- **Muestras aleatorias:** los individuos de la muestra son seleccionados AL AZAR (por ejemplo, numeramos la población y tomamos números al azar). Se corre el riesgo de que "el azar" provoque que la muestra no sea representativa.
- **Muestras intencionales:** el encuestador escoge a los individuos a los que estudiará. El problema es que la subjetividad del encuestador puede falsear el estudio.

Para saber más...

Aprende más sobre cómo seleccionar la muestra adecuada visitando el enlace "***Técnicas de muestreo***" que encontrarás en el apartado de recursos audiovisuales del tema.

¿Cómo obtengo los datos?

Una vez seleccionada la muestra, hay dos formas de **obtener la información** que necesitamos:

a) Obtención indirecta: los datos están recogidos y se consultan:

La **ventaja** es que ¡casi todo el trabajo está hecho! **Pero...**

- Puede que los datos no estén actualizados (¿y si el INE usase información de 1985 para el estudio que hemos visto?).
- Puede que los datos no se ajusten a nuestro estudio. Si queremos estudiar el tiempo que un grupo de niños dedican a desayunar y los datos disponibles se refieren a adultos ¡de poco nos sirven!

b) Obtención directa: los datos se observan **directamente sobre los individuos**, específicamente para el estudio. Seguramente los datos se ajustarán a nuestros deseos pero costará más obtenerlos.

- A veces se hará midiendo la característica en los individuos (por ejemplo, la estatura...)
- A veces se hará mediante cuestionarios que los encuestados responden...

Debemos tener en cuenta que…

Las preguntas de nuestra encuesta pueden influir en nuestra investigación.

Por ejemplo…

¿Son iguales estas dos preguntas para un cuestionario?

1. ¿Qué prefiere desayunar, yogurt desnatado o churros con chocolate?

2. ¿Qué suele desayunar con mayor frecuencia?

Es evidente que **las dos preguntas causarán un efecto muy diferente en el entrevistado**. Con la primera las opciones del encuestado son muy reducidas (es bastante probable que el individuo hubiese dado otra respuesta) y se ve claramente una intención oculta: separar falsamente a la población en dos grupos ficticios que no se corresponden con la realidad.

Por tanto es preferible la segunda pregunta, aunque con ella se corre un riesgo: puede que las respuestas obtenidas sean demasiado diferentes entre sí y la información sea poco útil... ¿y si todos responden una cosa diferente? ¿Podríamos extraer alguna conclusión?

Podemos obtener la información por…

observación directa o indirecta.

¿Qué hacemos con los datos que hemos tomado?

Vamos a suponer el siguiente estudio estadístico: queremos saber el tiempo que nuestros vecinos dedican a desayunar. Para ello hacemos un sorteo y **seleccionamos** para nuestra muestra **10 vecinos al azar de los 50** que hay en la vecindad y... ¡Les preguntamos...claro!

En este estudio tenemos:

- Variable aleatoria: X = tiempo dedicado al desayuno (en minutos). Es una variable cuantitativa discreta.
- N = 50, n = 10.
- Los individuos encuestados han dado las respuestas:

0, 0, 9, 0, 5, 5, 9, 5, 15, 0

Pero esta información no nos sirve demasiado. Vemos que se han dado los siguientes cuatro valores de la variable: x_1**=0,** x_2**=5,** x_3**=9,** x_4**=15**. ¿Se repite alguno más que otro? ¿Y si estudiásemos cuántas veces nos han respondido cada valor?

Nota: en matemáticas para representar un elemento en general que pertenezca a un conjunto de elementos ordenados, se usa el subíndice i. Así una valor genérico del conjunto $x_1, x_2, \ldots, x_4$, **se representará como** x_i, donde el subíndice i puede tomar los valores 1, 2,..., 4 (o los que haya, claro)

La frecuencia absoluta de un valor x_i de la variable es el número de veces que se ha observado dicho valor, y se representa n_i.

Por ejemplo, la frecuencia absoluta del valor x_1=0 es 4 (cuatro vecinos han respondido 0 minutos). Es decir $\mathbf{n_1=4}$. (Sólo hay que saber contar...)

Comprueba que lo has entendido

5. Ahora indica tú la frecuencia absoluta del resto de los valores de la variable:

   Recuerda, sólo hay que contar las veces que se repite cada valor, las veces que nos han respondido cada cantidad de minutos.

n_1 =	
n_2 =	
n_3 =	
n_4 =	
Y todas las frecuencias absolutas deben sumar...	

Vamos encaminados a elaborar una **TABLA DE FRECUENCIAS**, donde aparecerá **la información recogida de manera ORGNIZADA**. El siguiente paso es hallar otro tipo de frecuencias.

Hemos visto que 4 de los 10 vecinos no desayunan, casi la mitad. ¿Y si fuesen 4 vecinos de 4.566.123? ¡Prácticamente todos desayunarían! ¿A que la cosa cambia?

Las frecuencias absolutas **hay que "relativizarlas"...**

La frecuencia relativa de un valor de la variable es su frecuencia absoluta dividida por el número de observaciones. Para el valor x_i se representa f_i.

Así, como $n_1 = 4$, entonces $\mathbf{f_1 = 4:10 = 0.4}$ (la frecuencia absoluta del valor x_1=0 dividida entre 10 los vecinos encuestados).

Comprueba que lo has entendido

6. Ahora indica tú la frecuencia relativa de cada valor de la variable:

f_1 =	
f_2 =	
f_3 =	
f_4 =	
Y todas las frecuencias relativas deben sumar...	

Y todavía puede ser útil usar más tipos de frecuencias. . .

Tenemos las frecuencias absolutas (y relativas) acumuladas **para las que tenemos que ordenar los valores** de menor a mayor:

La frecuencia absoluta (o relativa) acumulada de un valor es la suma de todas las frecuencias absolutas (o relativas) de todos los valores MENORES O IGUALES QUE DICHO VALOR.

Se representan como las anteriores pero con mayúscula.

Por ejemplo, la frecuencia absoluta acumulada del valor 9 minutos es la suma de las frecuencias absolutas de los valores 0, 5 y 9.

Comprueba que lo has entendido

7. ¿Serías capaz tu solito/a de calcular las frecuencias absolutas y relativas ACUMULADAS de los valores de nuestra variable? Seguro que sí.

Frecuencias absolutas acumuladas		Frecuencias relativas acumuladas	
f_1 =		f_1 =	
f_2 =		f_2 =	
f_3 =		f_3 =	
f_4 =		f_4 =	

Y todos estos numeracos ¿para qué los hacemos? Pues con un único objetivo: para tener los datos bien ordenaditos, porque con todos estos números podemos **hacer la TABLA DE FRECUENCIAS** de nuestra variable.

Para hacer una tabla de frecuencias . . .

Se pone una primera columna con los valores de la variable y después una columna con cada tipo de frecuencia (y la suma o total en las no acumuladas), en nuestro caso:

Valores de la variables (minutos dedicados): x_i	Frecuencias absolutas: n_i	Frecuencias relativas: f_i	Frecuencias absolutas acumuladas: N_i	Frecuencias relativas acumuladas: F_i
$x_1 = 0$	$n_1 = 4$	$f_1 = 0,4$	$N_1 = 4$	$F_1 = 0,4$
$x_2 = 5$	$n_2 = 3$	$f_2 = 0,3$	$N_2 = 7$	$F_2 = 0,7$
$x_3 = 9$	$n_3 = 2$	$f_3 = 0,2$	$N_3 = 9$	$F_3 = 0,9$
$x_4 = 15$	$n_4 = 1$	$f_4 = 0,1$	$N_4 = 10$	$F_4 = 1$
total:	10	1		

Aunque, utilizando los símbolos con los que se representa cada cosa, basta escribir esto:

x_i	n_i	f_i	N_i	F_i
0	4	0,4	4	0,4
5	3	0,3	7	0,7
9	2	0,2	9	0,9
15	1	0,1	10	1
total:	10	1		

Comprueba que lo has entendido

8. Extrae alguna conclusión de la tabla anterior.
9. Realiza la tabla de frecuencias asociada al siguiente estudio...
   - X = nº de piezas de fruta consumidas al día.
   - P = habitantes de Nofrutataun.
   - N = 500
   - n = 25

   Respuestas: 2, 3, 4, 2, 1, 2, 0, 0, 3, 1, 2, 0, 1, 1, 0, 2, 0, 1, 4, 2, 3, 0, 1, 1, 2

¿Cómo podemos ver los datos gráficamente?

Los datos estadísticos se pueden "ver" en dibujos. Sí sí, en serio. ¿A que si lo logramos será genial? Vamos allá.

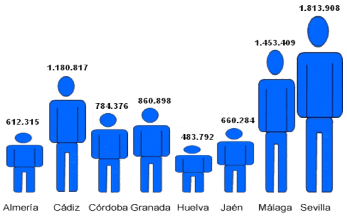
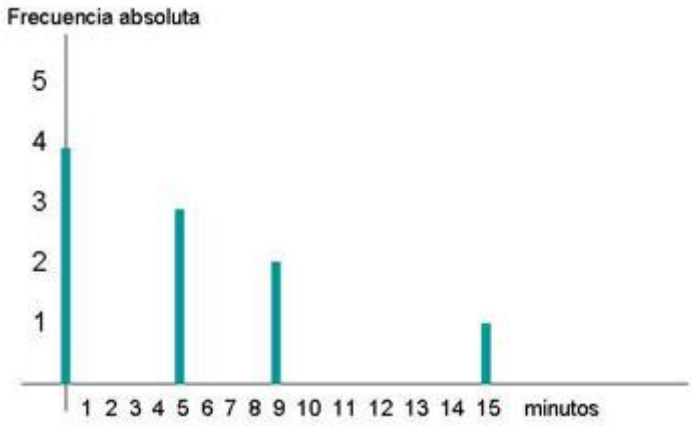


Diagrama de Barras:

Un diagrama de barras de una variable estadística se hace de forma muy sencilla. Por pasos, tras hacer la tabla de frecuencias:

- Primero dibujamos **dos ejes** de coordenadas.
- Después colocamos los **valores de la variable en el eje x**. Deben ser valores de variable discreta.
- Por último levantamos una **barra sobre cada valor**. ¿Hasta qué altura? Hasta lo que indiquen **las frecuencias** (absolutas o relativas).

Por ejemplo, en el caso anterior, a partir de la tabla podemos obtener el diagrama de barras de la imagen.



Un diagrama de barras se construye dibujando barras sobre los valores, que midan lo que indiquen las frecuencias absolutas (o relativas).

Diagrama de sectores:

Un diagrama de sectores es un círculo dividido en sectores cuya área será tanto mayor cuanto mayor sea la frecuencia del valor que representa.

- Dibujamos un **círculo**.
- Lo dividimos en tantos **sectores circulares** ("quesitos") como valores queramos representar. Pero claro, al valor con mayor frecuencia le corresponderá una parte más grande ¿verdad?
- Para cada valor su sector tendrá **un ángulo PROPORCIONAL** a su frecuencia. ¿Qué te has asustado? Vamos a ver cómo hacerlo que es fácil, sólo hay que hacer **una "regla de tres"**. El círculo abarca 360º ¿verdad? pues bien:

Valores	Frecuencia
0	4
5	3
9	2
15	1

  - Repartimos los 360º entre el total de respuestas que, en nuestro ejemplo, es 10.

    360º/10 = **36º grados para cada respuesta**.

  - Si cada respuesta se representa con un sector ("quesito") de 36º el valor 0 que se repitió **en cuatro respuestas** se representará con un sector de **4 · 36 = 144º**. Y así con el resto de valores.

Haciendo esto, obtenemos el diagrama de sectores de la imagen superior**.**

El cálculo que hemos tenido que hacer es: 360º/10 = 36º (Para ver cuántos grados corresponden a cada respuesta individual)

36·4 = 144º; 36·3 = 108º; 36·2 = 72º; 36·1 = 36º (Para ver cuántos grados le corresponde a cada dato según su frecuencia).

Para saber el ángulo que tiene que ocupar cada sector, dividimos 360º entre el número de respuestas y multiplicamos lo obtenido por la frecuencia absoluta de cada valor.

Pictograma

En este caso **representamos cada valor con un dibujo alusivo** cuyo TAMAÑO dependerá de la frecuencia absoluta (o relativa) del valor.

Por ejemplo, ya que estamos con el tiempo dedicado a desayunar, representaremos cada valor con una taza con reloj, que será más grande para las respuestas más frecuentes y menor para las respuestas menos frecuentes.

Pero cuidado

Si con tamaño nos referimos al área, entonces a doble frecuencia no corresponde doble tamaño, si no 4 veces el tamaño, porque la relación es cuadrática (un área se calcula elevando a cuadrado una longitud) y el dibujo obtenido no se corresponderá con los datos.

Por tanto, tenemos que dejar claro a qué nos referimos con tamaño. Podemos quedar en que a doble, triple... frecuencia corresponda doble, triple... altura del dibujo, por ejemplo.

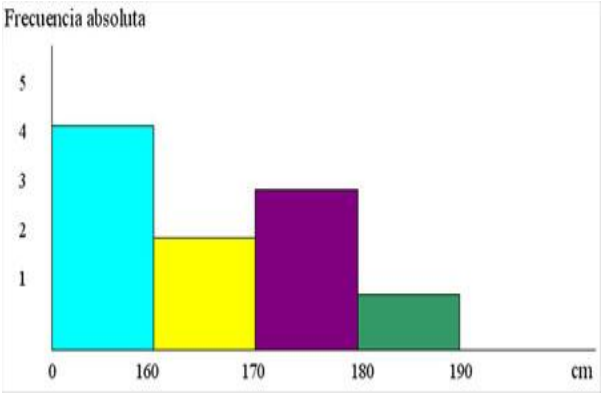
Histograma

Un histograma es muy **similar a un diagrama de barras**, solo que se usa **para variables cuyos datos se agrupan en intervalos**. Por ejemplo, si preguntamos a nuestros vecinos por su estatura, la tabla de frecuencias puede ser esta:

Valores de la variable (estatura en cm) I_i	Frecuencias absolutas n_i	Frecuencias relativas f_i	Frecuencias absolutas acumuladas N_i	Frecuencias relativas acumuladas F_i
150-160	4	0,4	4	0,4
160-170	2	0,2	6	0,7
170-180	3	0,3	9	0,9
180-190	1	0,1	10	1
	10	1		

¿Has visto algo nuevo en esta tabla? Seguro que te has dado cuenta de que en la primera columna no aparecen datos concretos de altura, sino los intervalos en que agrupamos los valores y las frecuencias absolutas se refieren al número de personas cuya altura está dentro del intervalo correspondiente.

Pues bien, para hacer el histograma:



- Primero dibujamos dos **ejes de coordenadas**.
- Después colocamos **los intervalos** en el eje X.
- Por último **levantamos un rectángulo sobre cada intervalo**. ¿Hasta qué altura? Pueden ocurrir dos cosas:
  - que todos los intervalos tengan la **misma amplitud** (como en nuestro caso, que todos "van de 10 en 10")
  - que haya intervalos más amplios que otros. Este caso lo dejaremos de momento.
- En el primer caso, la altura del rectángulo será **la indicada por la frecuencia** (absoluta o relativa) correspondiente.

Para saber más...

Los intervalos en los que se agrupan los valores de una variable continua o discreta con demasiados valores se llaman **intervalos de clase**.

El valor representante de un intervalo de clase se llama **marca de clase** y suele coincidir con el valor medio del intervalo (con el dato que queda justo en medio del mismo).

Interesantes ejemplos de gráficos

En el apartado de recursos del tema tienes una presentación con muchos gráficos estadísticos. Es una encuesta sobre salud nacional realizada en 2006 (Ministerio de Sanidad). ¡No tienes que aprenderte ninguno...! Son solo ejemplos para que veas cómo los gráficos estadísticos se usan con mucha "frecuencia".

Comprueba que lo has entendido

10. Realiza los gráficos estadísticos que puedas para la variable trabajada en el ejercicio anterior, el de las piezas de fruta. Te recordamos la tabla de frecuencias:

x_i	n_i	f_i	N_i	F_i
0	6	0.24	6	0.24
1	7	0.28	13	0.52
2	7	0.28	20	0.8
3	3	0.12	23	0.92
4	2	0.08	25	1
total	25	1		

Calculando números que informan sobre los datos

Antes de continuar, te vamos a preguntar una cosa. Si sacas en dos exámenes un 8 y un 6, ¿a que sabes tu nota media? Lo que haces, quizá sin saberlo, es sumar las dos notas y dividir lo obtenido por dos ¿verdad? Eso que haces se llama media aritmética.

Te vamos a contar más sobre ella y sobre alguna otra cosa... y haremos algunas cuentas.

La Media aritmética

No tiene nada que ver con  ni con 

Se llama media aritmética de una variable aleatoria a la suma de todos los valores observados dividida por el total de observaciones.

Volvamos al ejemplo del desayuno:

Hay **dos formas de calcular la media aritmética**. Una "a lo bruto" y otra "pensando un poco":

A lo bruto: sumamos todas las respuestas que nos han dado los vecinos y dividimos entre los 10 vecinos encuestados:

(0 + 0 + 9 + 0 + 5 + 5 + 9 + 5 + 15 + 0) :10 = 48

48:10 = 4,8

Esto quiere decir que si todos los vecinos desayunasen el mismo tiempo, **desayunarían todos 4,8 minutos**.

Pensando un poco: ¿y si en vez de sumar los valores como antes primero multiplicamos cada valor por las veces que se repite? Es decir: en la suma anterior hay, por ejemplo, tres cincos... ¿no sería mejor poner 15? Y así con el resto.

En realidad lo que haríamos sería **sumar cada valor multiplicado por su frecuencia absoluta y dividir después por el número total de observaciones**:

(0·4 + 5·3 + 9·2 + 15·1):10 = 48:10 = 4,8 minutos.

Podemos usar la tabla de frecuencias para este cálculo.

Ahora piensa: aunque hay un vecino que tarda en desayunar 15 minutos... ¿crees que los vecinos pasan bastante tiempo desayunando?

¿Qué ocurriría si. . .?

¿Qué ocurriría con la media anterior si un vecino tardase dos horas en tomarse su café y sus tostadas?

A la suma anterior tendríamos que añadir 120 minutos y obtendríamos 168, ahora dividimos entre el número de vecinos encuestados (que serían 11) y... SALE UNA MEDIA DE MÁS DE 15 MINUTOS.

¡Ya quisieran los demás tener tanto tiempo para desayunar! Se ha triplicado la media anterior.

Como ves, los "**valores extraños**" pueden producir **medias extrañas que no reflejen realmente la realidad**.

Para terminar ¿se te ocurre cómo calcular la media aritmética **cuando la variable está agrupada** en intervalos? Es fácil ¿no? Pues en vez de los valores **tomamos las marcas de clase** ¿recuerdas lo que eran?

Un caso especial:

¿Cuánto vale la media si todas las respuestas son iguales? Piensa un poco.

Por ejemplo, si todos los vecinos hacen el mismo número de comidas diarias (4), la media de comidas diarias es justo eso, 4 comidas.

¡Haz la cuenta y verás que no es magia!

La Moda

¿Qué quiere decir que un color está de moda? Pues que la mayoría de la gente se viste con ese color. ¿Qué significa que está de moda desayunar cereales? Que la mayoría de la gente toma cereales....

Y así podemos seguir.

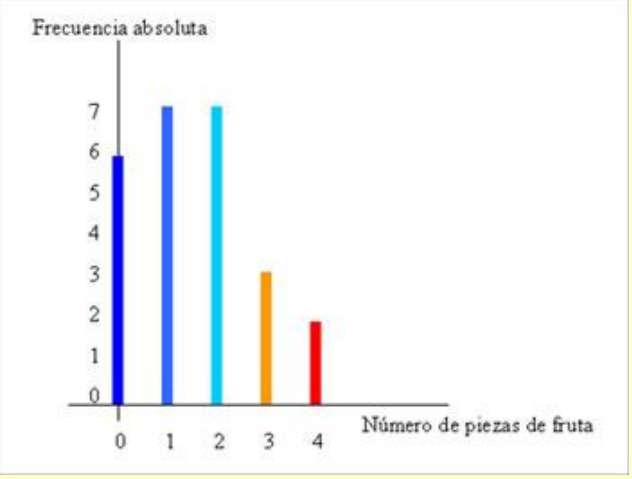
¿Qué será la moda de una variable estadística?

Efectivamente, acertaste

La moda de una variable estadística es el valor más frecuente, el más repetido en las respuestas... el de mayor frecuencia absoluta (o relativa).

Comprueba que lo has entendido

11. ¿Cuál es la moda en nuestro estudio del tiempo dedicado a desayunar?



Piensa un momento

¿Crees que puede haber **más de una moda** en una variable estadística?

Te ayudamos a pensarlo con una pequeña pista: sacude tu memoria... ¿te suena el gráfico de la derecha?

Efectivamente, puede haberla. Observa en la gráfica que hay dos valores con frecuencias iguales y ningún otro valor es más frecuente; ambos serán moda. Así podemos tener variables unimodales (con una moda), bimodales (con dos modas, como en el gráfico)...y hasta polimodales o multimodales (con **varias modas**).

Si la variable está agrupada en intervalos, podríamos indicar el **intervalo modal** (el intervalo que más datos "contiene").

Comprueba que lo has entendido

12. Calcula la media aritmética y la moda para la variable trabajada en los anteriores ejercicios, la de las piezas de fruta. Te recordamos la tabla de frecuencias:

x_i	n_i	f_i	N_i	F_i
0	6	0.24	6	0.24
1	7	0.28	13	0.52
2	7	0.28	20	0.8
4	2	0.08	25	1
total	25	1		

13. Halla la moda y el intervalo modal de la variable vista anteriormente relativa a la estatura.

    La tabla de frecuencias era…

Valores de la variable (estatura en cm) I_i	Frecuencias absolutas n_i	Frecuencias relativas f_i	Frecuencias absolutas acumuladas N_i	Frecuencias relativas acumuladas F_i
150-160	4	0,4	4	0,4
160-170	2	0,2	6	0,7
170-180	3	0,3	9	0,9
180-190	1	0,1	10	1
	10	1		

Para saber más...

La media se llama aritmética por una razón... y es que hay otras medias que no son aritméticas. Puedes verlo en el apartado de recursos web del tema, en el enlace: "***Otras medias y parámetros***".

Todos los datos no son iguales: dispersión de los datos

Observa los datos recogidos en estos dos estudios estadísticos:

Variable1:

NIVEL DE SATISFACCIÓN CON LA IMAGEN CORPORAL EN CIERTA CIUDAD

- **Población:** habitantes de Megusto con edades entre 11 y 40 años.
- **Muestra:** 230 habitantes de diversas edades, estamentos sociales y profesiones.
- **Pregunta:** ¿Qué nivel de satisfacción con su imagen tiene usted (de 0 a10)?

Nivel de satisfacción	Frecuencia absoluta
0	56
1	20
2	15
3	2
4	1
5	76
6	39
7	2
8	6
9	3
10	¿?

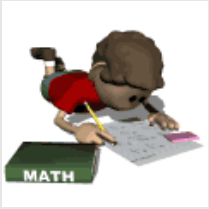
Influencia de los cánones de belleza	Frecuencia absoluta
0	1
1	2
2	1
3	2
4	2
5	5
6	9
7	20
8	45
9	56
10	87

Variable2:

INFLUENCIA DE LOS CÁNONES DE BELLEZA IMPUESTOS POR LASOCIEDAD COMO CAUSA DE LA ANOREXIA Y/O LA BULIMIA

- **Población:** habitantes de Nocomono con edades entre 11 y 40 años que han padecido o padecen trastornos alimenticios.
- **Muestra:** 230 habitantes de diversas edades y distintos estamentos sociales y profesiones que padecen o han padecido anorexia o bulimia.
- **Pregunta:** ¿Del cero al diez, cómo califica la influencia de los cánones de belleza impuestos por la sociedad como causa de su trastorno?

Trabajemos con ellos. . .



14. ¿Qué valor falta en la tabla primera?
15. Para ambos estudios, haz un diagrama de barras y halla la media aritmética y la moda.

(Al final del tema tienes las soluciones… para que compruebes qué tal lo has hecho)

Después de hacer las actividades anteriores… ¿Crees que en ambos casos las medias aritméticas son igual de representativas? La verdad es que no... Fíjate bien en los diagramas de barras y observa que los datos no están "distribuidos" de la misma manera en los dos ejemplos.

Ahora vas a estudiar que es posible expresar matemáticamente eso de "lo distribuidos que están los datos". Los tres números que vas a aprender a calcular (la varianza, la deviación típica y el coeficiente de variación) **indican lo agrupados o alejados que están los valores respecto de la media**, y por tanto **qué media da información más fiable sobre los datos.**

Varianza y desviación típica

Podemos estudiar si los datos de nuestras variables 1 y 2 están, en su conjunto, más o menos cerca de las medias respectivas.

Por ejemplo: En la variable1 quien ha respondido 10 se ha **alejado mucho de la media 3.51**. En cambio, los 12 que han respondido 3 o 4 se ha **acercado bastante a la media**. Cuanto más valores cercanos a la media tengamos, **más "fiable" es nuestra media** ¿verdad?

Existen dos números (parámetros estadísticos) que nos ayudan a **MEDIR esta cercanía de los datos a la media, es decir, a medir la dispersión de los datos**.

Uno es la **varianza**, que es la *media de las distancias de los valores a la media, al cuadrado*.

¿Qué te da miedo tanta palabrería...? Pues fíjate bien cómo no es para tanto...

Se hace así:

- Se calcula la distancia de cada valor a la media, sencillamente restándole al valor la media (para el valor 10 será 10 - 3,51= 6,49)
- Se elevan esas distancias al cuadrado (así, $6{,}49^2$=42,1201)
- Y se hace la media de los resultados como si fuesen valores (las frecuencias serán las originales)

Y si no te gusta el método anterior... ¡Hay otro mejor!

- Eleva los valores de la variable al cuadrado.
- Haz la media de los resultados obtenidos.
- Eleva la media de la variable al cuadrado y réstalo del resultado anterior.

Por ejemplo...

Para la variable 1...

- Los valores al cuadrado son:

  0^2 = **0,** 1^2 = **1,** 2^2 = **4,** 3^2 = **9,** 4^2 = **16,** 5^2 = **25,** 6^2 = **36,** 7^2 = **49,** 8^2 = **64,** 9^2 = **81,** 10^2 = **100.**

- La media de esos resultados es:

  (0x**56** + 1x**20** + 4x**15** + 9x**2** + 16x**10** + 25x**76** + 36x**39** + 49x**2** + 64x**6** + 81x**3** + 100x**1**) : 230

  (0+20+60+18+160+1900+1296+98+384+243+100):230 = 4279:230 = **18,604.**

- A esa cantidad restamos la media, 3,51, al cuadrado:

  18,604-3,51^2= 6.283.

Luego nuestra varianza es 6.283.

Muy muy importante

Como es una media de "números al cuadrado" y las cosas al cuadrado son siempre positivas, **LA VARIANZA ES SIEMPRE POSITIVA**.

Una varianza negativa se considera un "delito matemático", si te aparece alguna revisa tus cálculos porque te has equivocado seguro.

Si hemos calculado la varianza, la **desviación típica** es muy fácil de calcular: solo hay que **hacer la raíz cuadrada a la varianza**.

*(Por ejemplo: para la variable1 la desviación típica es la raíz cuadrada de 6,283, que es **2.506).***

Y **nos dice la dispersión respecto de la media**. ¿Y eso qué significa? Pues, en nuestro ejemplo, que los valores se alejan un promedio de 2.506 puntos respecto de la media, es decir, que muchos de los valores estarán entre 1 y 6 puntos, lo cual puede verse en la tabla y en el gráfico.

Comprueba que lo has entendido

16. El valor correcto de la varianza de la variable2 (con dos decimales) es…
17. El valor correcto de la desviación típica de la variable2 es…

Coeficiente de variación

Con los cálculos anteriores...

¿Podemos saber qué media de las dos anteriores es más "fiable"?

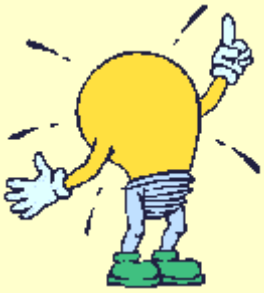
Las dos variables tratan cosas muy diferentes, y no podemos establecer la comparación. A priori puede parecer que la segunda variable tiene los datos MENOS DISPERSOS O MÁS AGRUPADOS, ya que su desviación típica es menor... ¿y si una variable estuviese medida en mm y la otra en número de sillas? Está claro que la comparación no es posible... ¿o sí?

Pues con los datos que hemos calculado no, pero sí con otro parámetro estadístico, es decir, otro "numerajo". Pero no te preocupes, **es el coeficiente de variación** y para calcularlo **basta dividir la desviación típica entre la media.**

En el ejemplo que estamos desarrollando...

El coeficiente de variación de la variable1 vale 2,506:3,51 = **0.7139**.

¿Para qué el coeficiente de variación?



Porque **el coeficiente de variación no tiene unidades** ya que se calcula dividiendo dos números que están en la misma unidad (desviación típica y media). Da igual que estemos hablando de sillas, mm o lo que sea.

Ahora sí podremos comparar los coeficientes y decidir que variable tiene la media más fiable…

A menor coeficiente de variación, menor dispersión, media más fiable.

Comprueba que lo has entendido

18. El valor correcto del coeficiente de variación de la variable2 es (con 2 decimales)…
19. ¿Qué datos están más cercanos a la media y, por tanto, qué media es más representativa y fiable?

Para saber más...

Las dos variables vistas están relacionadas con trastornos relativos a la alimentación. Para saber más sobre anorexia y bulimia puedes visitar los enlaces que encontrarás en el apartado de recursos web del tema:

- ***Geosalud***
- ***Té eres más que una imagen***
- ***Adaner***

El estudio estadístico se hace para sacar conclusiones

Ya sabes hacer muchas cosas con nuestros datos, pero no tiene sentido que nos pongamos a hacer cuentas y gráficos perdiéndonos entre números y tablas simplemente porque sí.

Ten presente que **se trata de dar respuesta** a lo que queríamos saber extrayendo conclusiones:

- **Las tablas** de frecuencias y los gráficos tienen por objetivo organizar y facilitar la visualización de los datos.
- **La media aritmética** indica un valor representativo de la variable, que resume la información de los datos recogidos y se interpreta en la realidad concreta que estamos estudiando.
- **La varianza y desviación típica** nos dicen cómo están de agrupados los datos respecto de la media. Permiten saber hasta qué punto la media aritmética da una buena información de la realidad estudiada.
- **El coeficiente de variación**, además de lo anterior, permite comparar la agrupación de los datos respecto de la media en distribuciones que no se parezcan en nada, lo cual puede resultar interesante.

Por ejemplo...

Imaginemos que estudiamos la dieta de 8 personas y extraigamos conclusiones a partir de los datos. Hemos estudiado dos variables: la *cantidad de calorías ingeridas por persona y día* y el *% de grasas ingeridas*.

Supongamos que, después de hacer todos los cálculos de nuestro estudio, estos son los resultados que hemos obtenido:

- **La media** de calorías ingeridas por persona y día son 1800 kcal, con una **moda** de 1810 kcal y una **desviación típica** de 50 kcal.
- El **porcentaje medio** de grasas que ingieren es un 40%, con una **moda** de 32% y una d**esviación típica** de 17 (17%).

¿Qué consecuencias podríamos sacar de estos resultados?...

Respecto a las calorías ingeridas, los valores de la media y la desviación típica nos dicen que l**a mayoría ingiere entre 1750 y 1850 kcal diarias, siendo el valor más repetido 1810 kcal**. Podemos interpretar que estas 8 personas tienen un **consumo de calorías diarias razonablemente sano** (acercándose a bajo), siendo su comportamiento muy parecido (puesto que hay poca variación entre las respuestas).

Respecto del % de grasas ingeridas, si bien no ingieren demasiadas calorías (como hemos visto antes) **el porcentaje medio de grasas en su alimentación es demasiado elevado**.

Sin embargo, la desviación típica indica que las respuestas están entre 23% y 57%, lo que refleja que hay grandes diferencias entre unas personas y otras. Algunas cuidan la cantidad de grasa en su dieta pero otras hacen un consumo peligroso para la salud.

Hay mucha disparidad en las respuestas:

	
Unos comen un número adecuado de calorías y una cantidad saludable de grasa. Seguramente hacen una dieta mediterránea, que es una de las mejores del mundo según los expertos... ¡Y la tenemos tan cerca!	Otros no comen demasiadas calorías, pero de lo que comen un porcentaje demasiado elevado es grasa. Pueden ser personas con poco tiempo para comer que abusan de la comida rápida, poco saludable.

Otro ejemplo...

- **Variable**: Índice de masa corporal (IMC) (para saber qué es el índice de masa corporal o IMC mira el enlace que encontrarás en el apartado de recursos web del tema).
- **Población**: niños de nuestro país.
- **Muestra**: 200 niños seleccionados al azar.

Tabla de frecuencias (parcial)	
IMC	**Frecuencia absoluta**
menor que 18,5	6
18,5 - 24,9	110
25-29,9	66
30-34,9	16
35-39,9	1
40 o mayor	1

¿Qué podemos decir a la vista de los datos? Seguro que se te ocurren muchas cosas, y eso que aún no hemos hecho ni una cuenta.

Casi a simple vista, vemos que…

- 110 niños (el **55%**) tienen un IMC **saludable**, entre 18,5 y 24,9.
- un **3%** tiene un peso demasiado **bajo**,
- y un **9% (18 de los 200) tienen obesidad.**

Además, el **IMC medio** es **24,51, es decir, normal cercano al sobrepeso.** (Hemos tomando como marcas de clase en los intervalos extremos la media de las 6 respuestas para el intervalo "menor que 18,5", que es 18 y LA respuesta en el intervalo "40 o mayor", que has sido 40).

Parece claro que todos tenemos que poner de nuestra parte, y las autoridades sanitarias, para cuidar la alimentación de los niños, ya que un niño obeso es un joven enfermo. Las prisas, el trabajo, la comodidad no pueden ganar la batalla a la dedicación y el interés por lo que nuestros niños comen.

Ya ves. Se pueden sacar muchas conclusiones, a veces de gran importancia social, con un estudio estadístico... ¡y eso que sólo hemos metido un pie en el mundo de la Estadística!

Para saber más...

¿Quieres información sobre cómo hacer una buena dieta y cuidar así tu salud y al de los tuyos? Visita el enlace sobre la dieta mediterránea que encontrarás en el apartado de recursos web del tema.

Comprueba que lo has entendido

20. Ahora extrae tú tus propias **conclusiones**, sin hacer nuevos cálculos, sobre:
    a. El estudio sobre el número de piezas de fruta.
    b. Las variables 1 y 2 que estudiaste en el apartado anterior.

Comprueba que lo has entendido (soluciones)

1. Son cuantitativas el número de horas de sueño, la estatura y el número de libros leídos el último mes. De estas variables, sólo la estatura es de tipo continuo.
2. Aunque podríamos idear muchos ejemplos, como variable cuantitativa podemos escoger el número de calorías consumidas en el desayuno, y como variable cualitativa el tipo de café bebido.
3. Claro que hay que tener cuidado con el número. Si seleccionamos demasiados no tendremos ninguna ventaja... ya puestos sería mejor tomar toda la población. Y si son demasiado pocos... quizá no sean representativos de la población y los datos no informen de lo que realmente ocurre.
4. Claro que sí, no es lo mismo preguntar hábitos de alimentación a deportistas que a diabéticos, por ejemplo.
5. La respuesta es:

n_1 =	4
n_2 =	3
n_3 =	2
n_4 =	1
y todas las frecuencias absolutas deben sumar…	10

6. La respuesta es:

f_1 =	0,4
f_2 =	0,3
f_3 =	0,2
f_4 =	0,1
Y todas las frecuencias relativas deben sumar…	1

7. La respuesta es:

	Frecuencias absolutas acumuladas		Frecuencias relativas acumuladas
f_1 =	4	f_1 =	0,4
f_2 =	7	f_2 =	0,7
f_3 =	9	f_3 =	0,9
f_4 =	10	f_4 =	1

Fíjate en que la última frecuencia absoluta acumulada coincide con el número de datos y que la última frecuencia relativa acumulada siempre es 1.

8. Algunas conclusiones que podríamos sacar son las siguientes:
   - El 40% (4 de 10 encuestados) de los vecinos NO desayuna. Es la respuesta más frecuente (repetida).
   - Sólo 1 de 10 dedica más de 10 minutos al desayuno (sólo el 10% de los vecinos).
   - 7 de los 10 vecinos encuestados dedica 5 minutos o menos a su desayuno (esto lo hace el 70% de los vecinos).
   - Por tanto, parece claro que el tiempo dedicado es por lo general insuficiente para un desayuno de calidad.

9. La tabla quedaría así:

x_i	n_i	f_i	N_i	F_i
0	6	0.24	6	0.24
1	7	0.28	13	0.52
2	7	0.28	20	0.8
3	3	0.12	23	0.92
4	2	0.08	25	1
total	25	1		

10. Como los valores no están agrupados en intervalos, podríamos hacer los gráficos siguientes:

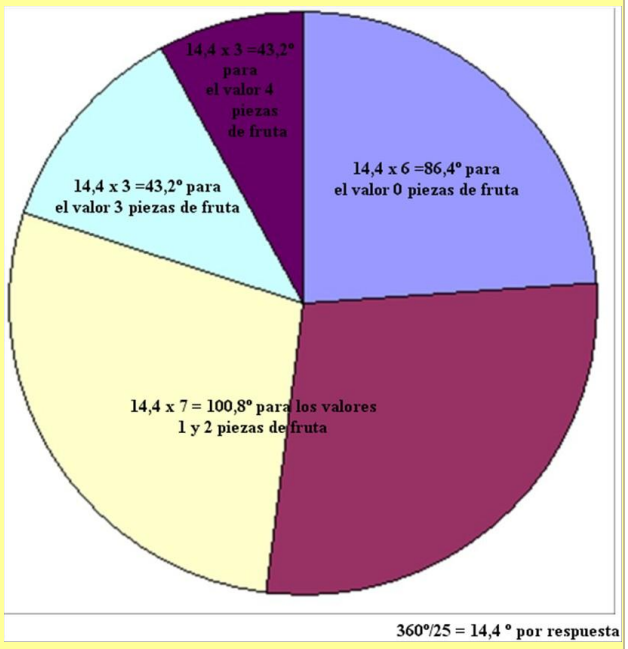
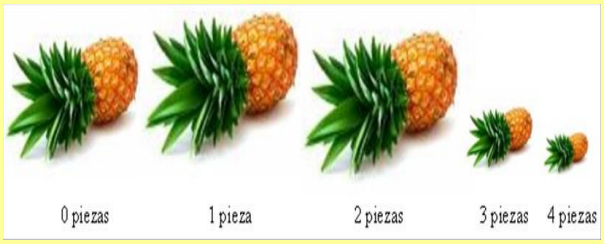


DIAGRAMA DE SECTORES



PICTOGRAMA

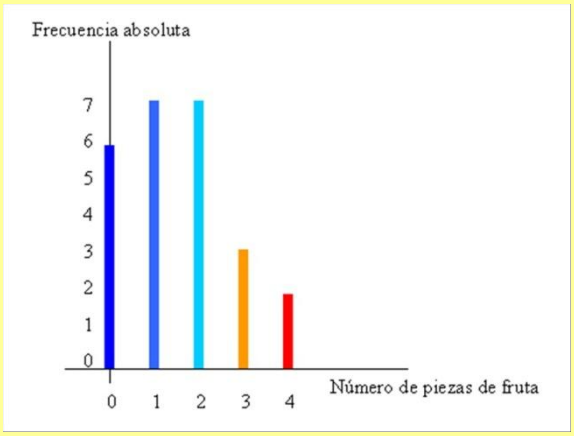


DIAGRAMA DE BARRAS

11. De las distintas respuestas una se repitió más que las demás: 5 minutos… Pues esa será la moda en nuestro estudio.
12. Media aritmética es: 1,52 piezas de fruta y la moda es 1 y 2 piezas (es una variable bimodal)
13. Media aritmética es: 166 cm y el intervalo modal es 150 – 160 cm.
14. Para saber la frecuencia que falta sólo hay que recordar que si las sumamos TODAS debe salirnos 230 (el total de observaciones). Como las demás suman 229, sólo una persona ("la número 230") responde 10.

15. Las soluciones son…

Variable 1:

- **La media** vale :

(0·**56**+1·**20**+2·**15**+3·**2**+4·**10**+5·**76**+6·**39**+7·**2**+8·**6**+9·**3**+10·**1**)/230

(0+20+30+6+40+380+234+14+48+27+10):230

809:230 = **3,51.**

- **La moda** es **5**, cuya frecuencia 76 es la mayor.

Variable 2:

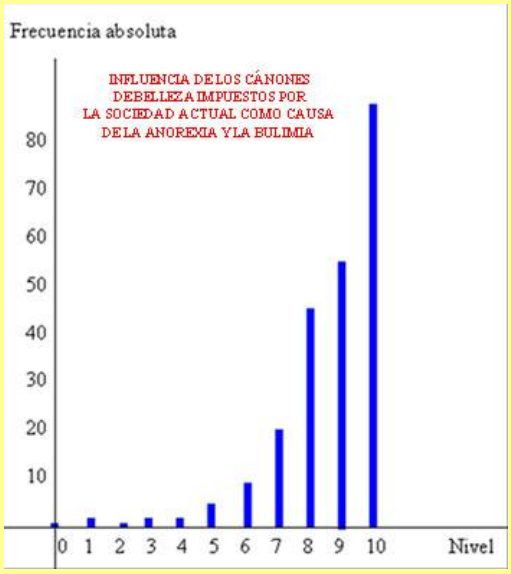
- **La media** vale:

(0·**1**+1·**2**+2·**1**+3·**2**+4·**2**+5·**5**+6·**9**+7·**20**+8·**45**+9·**56**+10·**87**)/230

(0+2+2+6+8+25+54+140+360+504+870):230

1971:230 = **8,56.**

- **La moda** es **10**, cuya frecuencia 87 es la mayor.



16. El valor correcto de la varianza de la variable2 (con dos decimales) es… 2,70.
17. El valor correcto de la desviación típica de la variable2 es… 1,64.
18. El valor correcto del coeficiente de variación de la variable2 es (con 2 decimales)… 0,19.
19. La media de la variable 2 es mucho más fiable que la de la variable 1, puesto que el coeficiente de variación es mucho más pequeño en la variable 2 que en la 1.
20. Esto de extraer conclusiones es siempre algo muy personal, pero hay algunas cosas que están más o menos claras:
    a. **Posibles conclusiones son:**
        - El número **medio** de piezas de fruta consumidas al día es **1,52**.
        - Las respuestas más frecuente son **1 y 2** piezas de fruta.
        - Más de la mitad (52%) toma **una o ninguna** pieza de fruta.
        - Sólo un 20% toma **3 o más** piezas.
        - Se debe promover un mayor consumo de frutas como parte de una alimentación completa y sana. Debemos sustituir por fruta la bollería industrial, perjudicial para nuestra salud y calidad de vida.

b. Veamos:

1. **Variable 1:**
   - Por término medio la puntuación de la imagen personal es **3,51**, lo que indica bajo nivel de autoestima respecto del aspecto físico.
   - La puntuación **más frecuente es 5**, indicada por 103 de los 230 encuestados.
   - Casi el **45% "suspende"** su imagen.
   - Sólo **el 0.52% está satisfecho** con su imagen (puntuándola con 7 o más).
   - Se deben promover patrones de belleza saludables y realistas, ya que los actuales producen frustración y baja autoestima respecto del aspecto físico.
2. **Variable 2:**
   - El nivel medio es **8,56**.
   - La respuesta **más frecuente es 10**.
   - Un elevado porcentaje (superior al **90%**, 208 de los 230 encuestados) atribuye una gran importancia como causa de su enfermedad (respondiendo 7 o más)
   - Sólo un **0.34%** atribuye un nivel de importancia **inferior a 5**.
   - Parece claro que la influencia de los cánones de belleza promovidos socialmente es un factor importante en la aparición de trastornos alimentarios, por tanto todos los componentes de la sociedad deberían tomar medidas al respecto, para que disminuyan los casos de éstas enfermedades.